ENGLISH

TRANSLATION OF

INTERNATIONAL APPLICATION

AS FILED

Description

METHOD FOR DETECTING TARGET SOUND, METHOD FOR DETECTING DELAY TIME IN SIGNAL INPUT, AND SOUND SIGNAL PROCESSOR

Technical Field

The present invention relates to a method for detecting a target sound and a program therefor, a method for detecting a delay time in signal input between sound signals inputted into plural microphones and a program therefor, a sound signal processor for processing inputted sound signals, and a voice recognition device for detecting a speech sound and processing voice recognition of the speech sound.

Background Art

In various forms of communication employed by humans, voice is the most basic and excellent communication means, with its information transmission speed higher than any other information transmission method. Thus, the voice has served as the basis of human communication means since ancient times until nowadays.

There are proposed voice recognition techniques for recognizing the voice. Voice recognition includes extracting the most basic information on the semantic contents, or phonological information, from the information contained in the voice with a computer or the like, and determining the extracted contents. In recent years, they have been attempting to apply such voice recognition techniques as a man-machine interface in various fields, with the drastic development of computer processor technology and the construction of advanced information networks, typically the Internet.

The recognition performance of current voice recognition systems has improved greatly by the utilization of probabilistic and statistical schemes. In the case of voice in ideal environments and voice collected at a short distance with a close-talking microphone, a significantly high recognition rate can be obtained.

However, when it comes to voice recognition in actual environments, the recognition rate is inferior, because of the mismatch between

learning data and observed data in their environments, contents of speeches, and the like. In addition, the users suffer great burden and discomfort from a close-talking microphone headset as a sound reception system worn by the user. This significantly hinders the practical application of voice recognition systems.

Further, many studies have been conducted on voice recognition methods using plural remote microphones for picking up remote voice, which is difficult to recognize owing to its lower S/N ratio, influences of background noise and room reverberation, and the like. A typical one of them is a method using a microphone array. This method can perform three types of spatial signal processing, namely sound source position detection processing, target sound emphasis processing, and noise suppression processing. Remote voice recognition is being extensively researched using methods such as the above.

However, this method requires plural microphones to be fixed at regular intervals for accurate identification processing of the direction of the speaker and thus the downsizing and mobilization is difficult. Therefore, there is a problem that this method is difficult to apply to voice input in various environments and under various circumstances and thus has limited uses.

As a "ubiquitous" sound reception system enabling anytime/anywhere sound input, there is an expectation of mountable microphones that can be attached to clothes, glasses or the like, which (1) are compact and lightweight for easy mounting/removing, (2) can ensure short-distance sound pickup generally as good as close-talking microphones, and (3) can ease the burden and discomfort when mounted to the user compared to close-talking microphone headsets.

The present invention has been made in view of the foregoing problems, and therefore has an object of providing a method for detecting a target sound, a method for detecting a delay time in signal input, a sound signal processor, a voice recognition device, and programs therefor, which enable the construction of a sound reception system employing plural mountable microphones and robust against environmental fluctuations.

Disclosure of the Invention

A method for detecting a target sound according to the present invention comprises: inputting detection target sounds outputted from a detection target sound source into plural microphones; detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; detecting an inclination of the phase of the cross-spectrum with respect to the frequency due to respective distances from the detection target sound source to the plural microphones; and, based on the inclination, detecting the target sound received by the plural microphones.

The above method for detecting a target sound may comprise: dividing the frequency according to the band; and, based on inclinations of the phase of each band divided, detecting the target sound.

The above method for detecting a target sound may comprise: detecting the target sound when a tendency that the inclinations of each band concentrate on a specific inclination is strong.

The above method for detecting a target sound may comprise: dividing the sound signals inputted into the plural microphones into predetermined time sections; and detecting the phase of the cross-spectrum between the sound signals in each section.

A method for detecting a delay time in signal input according to the present invention comprises: inputting sounds outputted from a sound source into plural microphones; detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; detecting an inclination of the phase of the cross-spectrum with respect to the frequency due to respective distances from the sound source to the plural microphones; and, based on the inclination, detecting the delay time in sound reception from the sound source between the plural microphones.

The above method for detecting a delay time in signal input may comprise: dividing the frequency according to the band; and, based on inclinations of the phase of each band divided, detecting the delay time in the sound reception.

The above method for detecting a delay time in signal input may comprise: detecting the delay time in the sound reception when a

tendency that the inclinations of each band concentrate on a specific inclination is strong.

The above method for detecting a delay time in signal input may comprise: dividing the sound signals inputted into the plural microphones into predetermined time sections; and detecting the phase of the cross-spectrum between the sound signals in each section.

A sound signal processor according to the present invention comprises: cross-spectrum phase detection means for detecting a phase of a cross-spectrum between sound signals inputted into plural microphones; inclination detection means for detecting an inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; and target sound detection means for detecting a target sound outputted from a detection target sound source and received by the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means.

The above sound signal processor may be characterized in that the inclination detection means divides the frequency of the phase of the cross-spectrum according to the band and detects inclinations of each band divided; and that the target sound detection means detects the target sound based on the inclinations of each band detected by the inclination detection means.

A sound signal processor for processing a sound outputted from a sound source and inputted into plural microphones according to the present invention comprises: cross-spectrum phase detection means for detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; inclination detection means for detecting an inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; delay time detection means for detecting a delay time in sound reception from the sound source between the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means; and sound signal synthesizing means for synthesizing the sound signals inputted into the plural microphones based on the delay time detected by the delay time detection

means.

The above sound signal processor may be characterized in that the inclination detection means divides the phase of the cross-spectrum according to the band and detects inclinations of each band divided; and that the delay time detection means detects the delay time in the sound reception based on the inclinations of each band detected by the inclination detection means.

A sound signal processor for processing a detection target sound outputted from a detection target sound source and inputted into plural microphones according to the present invention comprises: cross-spectrum phase detection means for detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; inclination detection means for detecting an inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; delay time detection means for detecting a delay time in sound reception from the detection target sound source between the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means; sound signal synthesizing means for synthesizing the sound signals inputted into the plural microphones based on the delay time detected by the delay time detection means; and target sound detection means for detecting the target sound in the synthesized sound signals synthesized by the sound signal synthesizing means based on the inclination with respect to the frequency detected by the inclination detection means.

The above sound signal processor may be characterized in that the inclination detection means divides the phase of the cross-spectrum according to the band and detects inclinations of each band divided; that the delay time detection means detects the delay time in the sound reception based on the inclinations of each band detected by the inclination detection means; and that the target sound detection means detects the target sound based on the inclinations of each band detected by the inclination detection means.

A voice recognition device for processing a speech sound outputted from a speech sound source and inputted into plural microphones

according to the present invention comprises: cross-spectrum phase detection means for detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; inclination detection means for detecting an inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; speech sound detection means for detecting the speech sound received by the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means; and voice recognition processing means for performing voice recognition processing of the speech sound detected by the speech sound detection means.

The above voice recognition device may be characterized in that the inclination detection means divides the frequency of the phase of the cross-spectrum according to the band and detects inclinations of each band divided; and that the speech sound detection means detects the speech sound based on the inclinations of each band detected by the inclination detection means.

A voice recognition device for processing a speech sound outputted from a speech sound source and inputted into plural microphones according to the present invention comprises: cross-spectrum phase detection means for detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; inclination detection means for detecting an inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; delay time detection means for detecting a delay time in sound reception from the speech sound source between the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means; sound signal synthesizing means for synthesizing the sound signals inputted into the plural microphones based on the delay time detected by the delay time detection means; speech sound detection means for detecting the speech sound in the synthesized sound signals synthesized by the sound signal synthesizing means based on the inclination with respect to the frequency detected by the inclination detection means; and voice recognition processing means for performing voice recognition

processing of the speech sound detected by the speech sound detection means.

The above voice recognition device may be characterized in that the inclination detection means divides the phase of the cross-spectrum according to the band and detects inclinations of each band divided; that the delay time detection means detects the delay time in the sound reception based on the inclinations of each band detected by the inclination detection means; and that the speech sound detection means detects the speech sound based on the inclinations of each band detected by the inclination detection means.

A program according to the present invention makes a computer perform processing of detecting a target sound, the processing comprising: inputting detection target sounds outputted from a detection target sound source into plural microphones; detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; detecting an inclination of the phase of the cross-spectrum with respect to the frequency due to respective distances from the detection target sound source to the plural microphones; and, based on the inclination, detecting the target sound outputted from the detection target sound source and received by the plural microphones.

A program according to the present invention makes a computer perform processing of detecting a delay time in sound reception, the processing comprising: inputting sounds outputted from a sound source into plural microphones; detecting a phase of a cross-spectrum between sound signals inputted into the plural microphones; detecting an inclination of the phase of the cross-spectrum with respect to the frequency due to respective distances from the sound source to the plural microphones; and, based on the inclination, detecting the delay time in sound reception from the sound source between the plural microphones.

Examining the phase of a cross-spectrum of plural sound signals picked up by plural microphones, the inclination of the phase with respect to the frequency is constant, depending on the difference between the respective distances from the sound source to the

microphones. The difference between the respective distances from the sound source to the microphones appears as a delay time in sound reception between the plural microphones. When the S/N ratio of the sound picked up by the plural microphones is higher, the tendency to such a constant inclination is more notable. The present invention utilizes this relation.

That is, in the present invention, the phase of a cross-spectrum between sound signals inputted into plural microphones is detected; the inclination of the phase of the cross-spectrum with respect to the frequency due to the respective distances from the sound source to the plural microphones is detected; and, based on the inclination, a detection target sound or speech sound received by the plural microphones is detected. The detection target sound includes sound produced by substances, in addition to speech sound produced by humans.

The present invention is based on the principle that, examining the phase of a cross-spectrum of plural sound signals picked up by plural microphones, the inclination of the phase with respect to the frequency is constant, depending on the difference between the distances from the sound source to the microphones, and that the tendency to such a constant inclination is more notable when the S/N of the sound picked up by the plural microphones is higher.

In the present invention, also, the phase of a cross-spectrum between sound signals inputted into plural microphones is detected; the inclination of the phase of the cross-spectrum with respect to the frequency due to the respective distances from the sound source to the plural microphones is detected; and, based on the inclination, a delay time in sound reception between the plural microphones is detected.

The present invention is based on the principle that, examining the phase of a cross-spectrum of plural sound signals picked up by plural microphones, the inclination of the phase with respect to the frequency is constant, depending on the difference between the respective distances from the sound source to the microphones, and that the difference between the respective distances from the sound source to the microphones appears as a delay time in sound reception between

the plural microphones.

In the present invention, the frequency of the phase of a cross-spectrum is divided according to the band, and the processing is performed based on the inclinations of each band divided. This allows detection of the inclinations with high accuracy.

Brief Description of Drawings

FIG. 1 is a block diagram showing the entire construction of a system including a sound signal processor of an embodiment of the present invention.

FIG. 2 is a block diagram showing the construction of a sound signal processor of a first embodiment of the present invention.

FIG. 3 is a property diagram showing the phase of a cross-spectrum in respective environments.

FIG. 4 is a property diagram showing the phase of a cross-spectrum, in which (A) is a property diagram showing the phase of a cross-spectrum of a voiced frame and (B) is a property diagram showing the phase of a cross-spectrum of a voiceless frame.

FIG. 5 is a property diagram showing a histogram obtained based on the phase of a cross-spectrum, in which (A) is a property diagram showing a histogram of a voiced frame and (B) is a property diagram showing a histogram of a voiceless frame.

FIG. 6 is a block diagram showing the construction of a histogram etc. calculating section and the like of the sound signal processor.

FIG. 7 is a property diagram used for describing the effects of the sound signal processor of the first embodiment.

FIG. 8 is a block diagram showing the construction of a sound signal processor of a second embodiment of the present invention.

FIG. 9 is a diagram used for describing the Overlap-add method for generating synthesized signals.

FIG. 10 is a property diagram used for describing the effects of the sound signal processor of the second embodiment.

FIG. 11 is a block diagram showing the construction of a sound signal processor of a third embodiment of the present invention.

FIG. 12 is a block diagram showing another construction of a

voiced/voiceless determining section of the sound signal processor.

Best Mode for Carrying Out the Invention

An embodiment of the present invention is described below in detail with reference to the drawings. As shown in FIG. 1, this embodiment is a sound signal processor 10 for processing sound signals picked up by two microphones 1 and 2. The first and second microphones 1 and 2 are of a mountable type that can be mounted to a sound source (user) with a comparatively high degree of freedom in their mounted positions.

FIG. 2 shows the construction of the sound signal processor 10 of a first embodiment. As shown in FIG. 2, the sound signal processor 10 includes first and second framing sections 11 and 12, first and second frequency analyzing sections 13 and 14, a cross-spectrum calculating section 15, a phase extraction processing section 16, a phase unwrap processing section 17, a main calculating section 30, and a sound input on/off control section 18. The main calculating section 30 includes a frequency band dividing section 31, first through N-th inclination calculating section 32_1 through 32_N, a histogram etc. calculating section 33, and a voiced/voiceless determining section 34. The processing operation of each section is described below.

Two-channel sound signals inputted from the first and second microphones 1 and 2 are inputted into the first and second framing sections 11 and 12, respectively. The sound signals inputted from the first microphone 1 are also inputted into the sound input on/off control section 18.

The first and second framing sections 11 and 12, the first and second frequency analyzing sections 13 and 14, and the cross-spectrum calculating section 15 calculate a cross-spectrum of the two-channel sound signals inputted from the first and second microphones 1 and 2.

For example, when sound signals picked up by plural microphones, such as the first and second microphones 1 and 2, are observed in a time series, there is a phase difference between the received sound signals. This results from the difference between the arrival times

of the sound signals from the sound source to the microphones 1 and 2 due to the difference between the distances from the sound source to the microphones 1 and 2.

Here, a case is examined in which: the delay time between the sound signals picked up by the first and second microphones 1 and 2 is measured; the phases of those signals are synchronized based on the measured delay time; and then the sound signals picked up by the first and second microphones 1 and 2 are added to obtain synchronized added sound. Such a technique for obtaining synchronized added sound as described above is disclosed in, for example, a literature "Acoustic event loocalization using a crosspower-spectruum phase based technique," by M. Omologo, P. Svaizer et al., Proc. ICASSP94, pp. 274-276 (1994).

The sound signals picked up by the two, first and second microphones 1 and 2 are represented as $x_1(t)$ and $x_2(t)$, respectively, and frequency functions obtained by Fourier transformations of these sound signals $x_1(t)$ and $x_2(t)$ are represented as $X_1(\omega)$ and $X_2(\omega)$, respectively. The sound signal $x_2(t)$ is assumed to be a time-shifted waveform of the sound signal $x_1(t)$ as represented by the following equation (1):

$$x_2(t) = x_1(t - t_0) \qquad (1)$$

On this assumption, the relation between the frequency functions $X_1(\omega)$ and $X_2(\omega)$ can be represented by the following equation (2):

$$X_2(\omega) = e^{-j\omega t0}X_1(\omega) \qquad (2)$$

Then, from the frequency functions $X_1(\omega)$ and $X_2(\omega)$, a cross-spectrum $G_{12}(\omega)$ can be obtained as represented by the following equation (3):

$$G_{12}(\omega) = X_1(\omega)X_2^*(\omega) = X_1(\omega)e^{j\omega t0}X_1^*(\omega) = |X_1|^2e^{j\omega t0} \qquad (3)$$

The exponent term of the cross-spectrum $G_{12}(\omega)$ corresponds to the time delay between the channels in the spectrum region. Thus, $X_2(\omega)e^{j\omega t0}$, obtained by multiplying the frequency function X_2 by the delay term $e^{j\omega t0}$, is synchronized with the frequency function X_1, whereby the inverse Fourier transform of $X_1(\omega) + X_2(\omega)e^{j\omega t0}$ can be dealt with as channel-synchronized-added sound.

The cross-spectrum $G_{12}(\omega)$ such as above is obtained by the cross-spectrum calculating section 15.

To this end, first of all, the first framing section 11 performs

framing of the sound signals inputted from the first microphone 1 (or divides them into frames), in preparation for the first frequency analyzing section 13 as the next step, and outputs the results to the first frequency analyzing section 13. Also, the second framing section 12 performs framing of the sound signals inputted from the second microphone 2 (or divides them into frames), in preparation for the second frequency analyzing section 14 as the next step, and outputs the results to the second frequency analyzing section 14. The first and second framing sections 11 and 12 progressively divide the inputted sound signals into frames, with each frame containing a predetermined number of samples.

For example, when no voice (speech) is inputted into the microphones 1 and 2, voiceless frames carrying no voice are generated, and when voice is inputted into the microphones 1 and 2, voiced frames carrying voice (speech) are generated.

The first frequency analyzing section 13 performs Fourier transformations of the sound signals from the first framing section 11 to calculate the frequency function $X_1(\omega)$, and outputs it to the cross-spectrum calculating section 15 as the next step. The second frequency analyzing section 14 performs Fourier transformations of the sound signals from the second framing section 12 to calculate the frequency function $X_2(\omega)$, and outputs it to the cross-spectrum calculating section 15 as the next step. The first and second frequency analyzing sections 13 and 14 perform a Fourier transformation for each frame of the sound signals.

The cross-spectrum calculating section 15 calculates the cross-spectrum $G_{12}(\omega)$ based on the frequency functions $X_1(\omega)$ and $X_2(\omega)$ obtained from the first and second frequency analyzing sections 13 and 14, using the equation (3).

FIG. 3 shows examples of the phase of a cross-spectrum of sound signals for one frame. In FIG. 3, (A) shows the phase of a cross-spectrum obtained from sound produced in a car, (B) shows the phase of a cross-spectrum obtained from sound produced in an office space, (C) shows the phase of a cross-spectrum obtained from sound produced in a soundproof room, and (D) shows the phase of a

cross-spectrum obtained from sound produced on a sidewalk (outdoor). As shown in FIG. 3, the phase of the cross-spectrum exhibits a generally constant inclination with respect to the frequency within a frame, in other words locally, depending on the difference between the distances from the sound source to the first and second microphones 1 and 2. In other words, the phase component of the cross-spectrum has a constant inclination depending on the difference between the distances from the sound source to the first and second microphones 1 and 2.

When the S/N ratio of the sound signals picked up by the first and second microphones 1 and 2 is higher, the tendency to such a constant inclination is more notable. Since the first and second microphones 1 and 2 are of a mountable type, the S/N ratio of the sound signals picked up by the first and second microphones 1 and 2 is high. Thus, each of the phases of the cross-spectra apparently exhibits a constant inclination.

The cross-spectrum calculating section 15 outputs a cross-spectrum $G_{12}(\omega)$ with such properties to the phase extracting section 16.

The phase extracting section 16 extracts (detects) the phase of the cross-spectrum $G_{12}(\omega)$ obtained from the cross-spectrum calculating section 15, and outputs the results of the extraction to the phase unwrap processing section 17.

The phase unwrap processing section 17 unwraps the cross-spectrum $G_{12}(\omega)$ based on the results of the phase extraction in the phase extracting section 16, and outputs the results of the upwrapping to the frequency band dividing section 31 of the main calculating section 30.

The frequency band dividing section 31 outputs segments obtained by dividing the phase according to the band to the first through N-th inclination calculating sections 32_1 through 32_N, respectively.

Note that there is a great difference in the phase components of a cross-spectrum between voiceless frames carrying no voice and voiced frames carrying voice. That is, the phase of a cross-spectrum has a generally constant inclination with respect to the frequency in voiced frames, whereas not in voiceless frames. A description is made

with reference to FIG. 4.

FIG. 4 shows examples of the phase of a cross-spectrum (CRS). In FIG. 4, (A) shows the phase of a cross-spectrum of a voiced frame, and (B) shows the phase of a cross-spectrum of a voiceless frame.

As can be seen from this comparison of FIG. 4(A) and FIG. 4(B), the phase of a cross-spectrum in voiceless frames has no specific trend with respect to the frequency. In other words, the phase of a cross-spectrum does not have a constant inclination with respect to the frequency. This is because the noise has a random phase.

On the other hand, the phase of a cross-spectrum in voiced frames has a constant inclination with respect to the frequency. This inclination depends on the difference between the distances from the sound source to the microphones 1 and 2.

As described above, there is a great difference in the phase components of a cross-spectrum between voiceless frames carrying no voice and voiced frames carrying voice.

In view of the above, the frequency band dividing section 31 divides the phase components into small frequency segments (or divides them according to the band) and the first through N-th inclination calculating sections 32_1 through 32_N as the next step calculate the inclinations of each segment by applying the least squares method, so as to follow the trend correctly even when the phase is rotated. The first through N-th inclination calculating sections 32_1 to 32_N respectively output the calculated inclination to the histogram etc. calculating section 33.

The method for obtaining the inclinations of each segment by applying the least squares method is a known technique disclosed, for example, in "Introduction to Signal Processing and Image Processing," by Nobukatsu Takai, Kougakusha (2000).

The histogram etc. calculating section 33 obtains a histogram based on the inclinations calculated by the first through N-th inclination calculating sections 32_1 to 32_N.

FIG. 5 shows histograms obtained by the histogram etc. calculating section 33, with each histogram showing inclinations by the segment. In other words, FIG. 5 shows the distribution of inclinations of the

phase, with the vertical axis representing the ratio, or incidence, of the segments of each inclination to all the segments. In FIG. 5, (A) shows a histogram of a voiced frame, and (B) shows a histogram of a voiceless frame.

As can be seen from this comparison of FIG. 5(A) and FIG. 5(B), in voiced frames, the histogram obviously has a peak value; that is, the inclinations are localized within a significantly narrow range, with a high incidence of inclinations of a specific range. In other words, the tendency that the inclinations of each band concentrate on a specific inclination is strong. On the other hand, in voiceless frames, the histogram takes a smooth shape, with the inclinations distributed over a wider range.

The histogram etc. calculating section 33 outputs the incidences obtained by creating these histograms to the voiced/voiceless determining section 34. A specific example of the processing performed by the histogram etc. calculating section 33 will be described later.

The voiced/voiceless determining section 34 determines voiced and voiceless sections based on the incidences obtained from the histogram etc. calculating section 33. For example, a section is determined to be a voiced section when the occurring incidence of inclinations included within a predetermined range around the mean value of the incidences is not less than a predetermined threshold, whereas a section is determined to be a voiceless section when that occurring incidence is less than the predetermined threshold.

Here, a frame is determined to be a voiced frame or a voiceless frame, since the processing at the previous step was performed by the frame. The voiced/voiceless determining section 34 outputs the determination results to the sound input on/off control section 18.

The sound input on/off control section 18 receives the sound signals from the first microphone 1, and switches on and off these sound signals to be outputted to the next step based on the determination results of the voiced/voiceless determining section 34. Specifically, when the voiced/voiceless determining section 34 determined sound signals to be a voiced section, the sound input on/off control section 18

switches on to output the sound signals to the next step. When the voiced/voiceless determining section 34 determined sound signals to be a voiceless section, the sound input on/off control section 18 switches off not to output the sound signals to the next step.

Here, the sound input on/off control section 18 switches on and off the part of the sound signals as the unit from the first microphone 1 corresponding to the frame on which the determination was made, since the processing at the previous step was performed by the frame.

A specific example of the processing performed by the histogram etc. calculating section 33 is described. FIG. 6 shows the construction of the histogram etc. calculating section 33 for implementing the processing.

The histogram etc. calculating section 33 includes a first switch 33S1, a second switch 33S2, and a mode calculating section 33C, as a construction for calculating an inclination of a high incidence (modal inclination) from the inclinations calculated by the first through N-th inclination calculating sections 32_1 through 32_N. The histogram etc. calculating section 33 switches on (closed) the first switch 33S1 for a given period, to create data (or a database) 33D1 of inclinations for the given period calculated by the first through N-th inclination calculating sections 32_1 through 32_N. Note that the second switch 33S2 is kept off (opened) at this time. When the data 33D1 are created, the second switch 33S2 is switched on (closed), to output the data 33D1 to the mode calculating section 33C.

The mode calculating section 33C creates a histogram representing the inclinations as shown in FIG. 5 from the data 33D1, and calculates the inclination of the highest incidence (hereinafter referred to as modal inclination) $\tau 0$ in the histogram. Instead of calculating the inclination of the highest incidence, it is also possible to calculate the inclination of the mean value $\tau 0$ or an inclination $\tau 0$ as a combination of the inclination of the highest incidence and the mean value of the inclinations. Thus, when the tendency that the inclinations of each band concentrate on a specific inclination is strong, the exact value, or an approximate value, of the specific inclination can be obtained. In this embodiment, the mode calculating

section 33C calculates the modal inclination $\tau 0$.

Then, the mode calculating section 33C outputs the calculated modal inclination $\tau 0$ to the voiced/voiceless determining section 34. The modal inclination $\tau 0$ is outputted to the voiced/voiceless determining section 34 as data 33D2.

The foregoing is one specific example of the processing performed by the histogram etc. calculating section 33.

The voiced/voiceless determining section 34 determines voiced and voiceless sections based on the modal inclination $\tau 0$ from the histogram etc. calculating section 33.

In the preceding description, the voiced/voiceless determining section 34 determined voiced and voiceless sections based on the incidences obtained from the histogram etc. calculating section 33. The voiced/voiceless determining section 34 determines voiced and voiceless sections based on the modal inclination $\tau 0$ obtained from the histogram etc. calculating section 33 and the inclinations (of each band) τi calculated by the first through N-th inclination calculating sections 32_1 through 32_N; therefore, the voiced/voiceless determining section 34 is adapted to receive the inclinations calculated by the first through N-th inclination calculating sections 32_1 through 32_N.

The voiced/voiceless determining section 34 compares the inclinations τi calculated by the first through N-th inclination calculating sections 32_1 through 32_N and the modal inclination $\tau 0$, using the following inequality (4):

$$|\tau i - \tau 0| < \delta \qquad (4)$$

wherein δ represents a threshold used for the determination (inclination threshold).

The voiced/voiceless determining section 34 determines a section to be a voiced section when the condition of the inequality (4) is satisfied with more than a predetermined ratio (YES), and determines a section to be a voiceless section when not (NO). Then, the voiced/voiceless determining section 34 outputs the determination results to the sound input on/off control section 18.

The sound signal processor 10 constructed as described above

functions consecutively as follows.

First of all, the first and second framing sections 11 and 12, the first and second frequency analyzing sections 13 and 14, and the cross-spectrum calculating section 15 calculate a cross-spectrum $G_{12}(\omega)$ of two-channel sound signals inputted from the first and second microphones 1 and 2.

Then, the phase extracting section 16, the phase unwrap processing section 17, and the frequency band dividing section 31 divide the phase of the thus calculated cross-spectrum $G_{12}(\omega)$ according to the band (divide them into segments), and the first through N-th inclination calculating sections 32_1 through 32_N calculate the inclinations of the phase of each band (each segment).

Then, the histogram etc. calculating section 33 generates a histogram based on the inclinations of each band (each segment) calculated respectively by the first through N-th inclination calculating sections 32_1 through 32_N, and the voiced/voiceless determining section 34 determines voiced and voiceless sections based on the incidences and the modal inclination $\tau 0$ obtained from the histogram. Based on the determination results, the sound input on/off control section 18 switches on and off the sound signals from the first microphone 1 to be outputted to the next step. Specifically, when the voiced/voiceless determining section 34 determined sound signals to be a voiced section, the sound input on/off control section 18 switches on to output the sound signals to the next step. When the voiced/voiceless determining section 34 determined sound signals to be a voiceless section, the sound input on/off control section 18 switches off not to output the sound signals to the next step.

In this manner, the sound signal processor 10 can detect speech sections (voiced sections) contained in the sound picked up by the first microphone 1 and 2.

Implementation of such a sound signal processor between the first microphone 1 and 2 and a voice application, for example, allows the voice application to securely perform processing related to speech sections. The voice application includes a voice recognition system, a broadcasting system, a cellular phone, and a transceiver. For

example, when the voice application is a voice recognition system, the voice recognition system can perform voice recognition based on the sound signals contained in speech sections outputted by the sound signal processor 10.

The effects are described next.

As described previously, the phase of a cross-spectrum between the sound signals inputted into the first and second microphones 1 and 2 is detected, and speech sections contained in the sound signals picked up by the plural microphones are detected based on the inclination of the detected phase of the cross-spectrum with respect to the frequency. In other words, speech sections contained in the sound signals picked up by the plural microphones are detected utilizing the great difference in the phase components of a cross-spectrum generated from sound signals containing no voice (speech) and sound signals containing voice (speech).

Specifically, the phase of the cross-spectrum is divided according to the band (divided into segments), a histogram is generated based on the inclinations of the phase of each band (each segment), an incidence (specifically mode) is obtained from the histogram, and speech sections are detected based on the incidence.

This allows accurate detection of speech sections. Further, utilizing such sound signals contained in the speech sections detected by the sound signal processor 10 allows voice recognition with a high recognition rate/low misrecognition rate in a voice recognition system, hands-free, half-duplex operation with high reliability in a cellular phone and a transceiver, and reduction of the power consumption of the communication system in a broadcasting system.

Even in the case of environmental changes, such as a change in the mounted positions of the microphones, and movement of the sound source, such as movement or a change in posture of the speaker, robust voice input can be achieved.

As described previously, the inclination of the phase of a cross-spectrum with respect to the frequency changes depending on the difference between the distances from the sound source to the first and second microphones 1 and 2. Thus, when the mounted positions of

the first and second microphones 1 and 2 relative to the sound source are changed, for example, the inclination of the phase of the cross-spectrum with respect to the frequency is also changed in response to the changes in the positions. Meanwhile, as describe previously, the phase of the cross-spectrum is divided according to the band (divided into segments), a histogram is generated based on the inclinations of the phase of each band (each segment), an incidence (specifically mode) is obtained from the histogram, and speech sections are detected based on the incidence. In other words, speech sections are detected eventually, irrespective of the magnitude of the inclination of the phase of the cross-spectrum, or the distances from the sound source to the microphones 1 and 2. Therefore, even when the mounted positions of the first and second microphones 1 and 2 relative to the sound source are changed, the detection results of speech sections are not affected.

As a result, even in the case of environmental changes, such as in the mounted positions of the microphones, and movement of the sound source, such as movement or a change in posture of the speaker, robust voice input can be achieved. In other words, robust voice input can be achieved while keeping a high degree of freedom in the positions of the microphones.

As described above, the aforementioned various effects can be attained even on the assumption of the use of mountable microphones, which are compact and lightweight for easy mounting/removing, can ensure short-distance sound pickup generally as good as close-talking microphones, and can ease the burden and discomfort when mounted to the user compared to close-talking microphone headsets.

(Example (First Embodiment))

A detection of a speech section containing voice was performed using a system to which the present invention was applied. Sample sound used was a total of forty sentences with a voiceless section of about one second intervening between sentences. Experiments were performed in the following environments: in a soundproof room, in a car, in an office space, and on a sidewalk. For evaluation, a frame was determined to be an error frame when (1) a voiceless frame was

misdetermined to be a voiced frame, or (2) judging from its leading end and trailing end, a speech section was determined to be a non-speech section. As a comparison object (conventional example), a method was used utilizing a Fisher's linear discriminant function using the average number of zero-crossings and the logarithmic power as variables.

FIG. 7 shows the results. FIG. 7 shows the percentage of the ratio of error frames to the total frames (speech section misdetection rate). In FIG. 7, the values designated as LDF are those obtained by the method utilizing the linear discriminant function, while the values designated as CRS are those obtained by the method utilizing the cross-spectrum (the present invention).

As shown in FIG. 7, in a soundproof room and in an office space, there was observed no great difference in resulting speech section misdetection rate between the method utilizing the average number of zero-crossings and the logarithmic power and the method of the present invention. However, in a car and on a sidewalk, the method of the present invention demonstrated improved results of the speech section misdetection rate. Thus, the present invention functions effectively particularly in noisy environments.

A second embodiment is described hereinafter.

FIG. 8 shows the construction of a sound signal processor 10 of the second embodiment. In the second embodiment, the sound signals picked up by the first and second microphones 1 and 2 are synthesized to be outputted to a voice application as the next step. To this end, the second embodiment includes a delay processing section 51 and a waveform synthesizing section 52. The delay processing section 51 delays the sound signals from the second microphone 2 and outputs them to the waveform synthesizing section 52, and the waveform synthesizing section 52 synthesizes the sound signals from the first microphone 1 and the sound signals of the second microphone 2 inputted from and delayed by the delay processing section 51 and outputs them.

There is observed a phase difference between the sound signals picked up by plural microphones, such as the first and second microphones 1 and 2, because of the difference between the distances from the sound

source to the microphones 1 and 2. Therefore, in order to synthesize the sound signals picked up by plural microphones such as the first and second microphones 1 and 2, the delay-and-sum processing is necessary, in which: the difference between the arrival times of the sound signals from the sound source to the microphones 1 and 2 is corrected; the phases of those signals are synchronized; and thereafter the sound signals are added. This is the reason that the second embodiment includes the delay processing section 51 and the waveform synthesizing section 52 as described previously.

In the foregoing first embodiment (see FIG. 6), the mode calculating section 33C calculated the modal inclination $\tau 0$ from the histogram. In the second embodiment, the delay processing section 51 performs delay processing based on the modal inclination $\tau 0$. A specific description is made below.

As shown in FIG. 3 and (A) of FIG. 4, the phase components of a cross-spectrum have a constant inclination in voiced sections. This inclination indicates the delay time between the channels of the first and second microphones 1 and 2.

Utilizing this relation, the delay processing section 51 performs delay processing based on the modal inclination $\tau 0$ calculated by the histogram etc. calculating section 33. Specifically, as shown in FIG. 6, the mode calculating section 33C outputs the modal inclination $\tau 0$ to the delay processing section 51, and the delay processing section 51 performs delay processing based on the inputted modal inclination $\tau 0$.

$$\tau 0 = x/n = 2\Pi \cdot n_0/N \ [\text{rad/point}] \qquad (5)$$

wherein the units for x and n are respectively radian and frequency point (point), N represents the number of FFT points, and n_0 represents the number of delay sampling points. From this relation, the number of delay sampling points n_0 using the modal inclination $\tau 0$ as a variable can be obtained by the following equation (6):

$$n_0 = \tau 0/(2\Pi/N) \ [\text{point}] \qquad (6)$$

Then, using this number of delay sampling points n_0, the delay time t_0 can be obtained by the following equation (7):

$$t_0 = n_0/F_S \qquad (7)$$

wherein F_s represents the sampling frequency, 16 kHz, for example.

The delay processing section 51 delays the sound signals inputted from the second microphone 2 based on the thus obtained delay time t_0, and outputs them to the waveform synthesizing section 52.

The waveform synthesizing section 52 synthesizes the sound signals from the first microphone 1 and the sound signals of the second microphone 2 inputted from and delayed by the delay processing section 51, and outputs them.

Synthesized sound signals may also be obtained in such a manner as described below.

As described previously, $X_2(\omega)e^{j\omega t0}$, obtained by multiplying the frequency function X_2 by the delay term $e^{j\omega t0}$, is synchronized with the frequency function X_1, whereby the inverse Fourier transform of $X_1(\omega) + X_2(\omega)e^{j\omega t0}$ can be dealt with as channel-synchronized-added sound. Utilizing this relation, synthesized sound signals are obtained.

That is, first of all, the delay time t_0 is used to obtain the channel-synchronized-added sound $X_1(\omega) + X_2(\omega)e^{j\omega t0}$ on the frequency scale by the following equation (8). Note that the delay time t_0 has the modal inclination $\tau 0$ as a variable as shown in the equations (6) and (7).

$$X_1(\omega) + X_2(\omega)e^{j\omega t0} = \{Re[X_1(\omega)] + jIm[X_1(\omega)]\} + \{Re[X_2(\omega)](\cos\omega t_0 + j\sin\omega t_0) + jIm[X_2(\omega)](\cos\omega t_0 + j\sin\omega t_0)\} \qquad (8)$$

Here, the channel-synchronized-added spectrum is a complex spectrum composed of a real part and an imaginary part represented respectively as follows:

$$\text{Re: } Re[X_2(\omega)]\cos\omega t_0 - Im[X_2(\omega)]\sin\omega t_0 + Re[X_1(\omega)]$$

$$\text{Im: } Re[X_2(\omega)]\sin\omega t_0 + Im[X_2(\omega)]\cos\omega t_0 + Re[X_1(\omega)]$$

This processing is performed for each frame and then IFFT (inverse FFT) is performed for each frame, to obtain a frame string of the synchronized added sound.

The Overlap-add method is then applied to the thus obtained frame string, to obtain synchronized added sound, or synthetic signals of the sound signals of the first microphone 1 and the sound signals of the second microphone 2.

The Overlap-add method is a method in which inputted data strings $s_n(t)$ are added in overlapping relation as shown in FIG. 9. Here, $s_n(t)$ represents an n-th synthesized sound waveform frame. The symbol L in the figure represents a constant.

In the sound signal processor 10 constructed as described above, the delay processing section 51 delays the sound signals from the second microphone 2 and outputs them to the waveform synthesizing section 52, and the waveform synthesizing section 52 synthesizes the sound signals from the first microphone 1 and the sound signals of the second microphone 2 inputted from and delayed by the delay processing section 51 and outputs them.

The effects achieved by this construction are as follows.

As described in connection with the foregoing first embodiment, the inclination of the phase of a cross-spectrum with respect to the frequency changes depending on the difference between the distances from the sound source to the first and second microphones 1 and 2. The delay time is estimated from this inclination of the phase of a cross-spectrum with respect to the frequency. The value actually used for the estimation is designated as modal inclination $\tau 0$. The use of the modal inclination $\tau 0$ in estimating the delay time ensures high accuracy of the estimated delay time.

Further, by synthesizing the sound signals of the first and second microphones based on the delay time as described above, there can be provided high-quality synthesized sound signals. For example, utilizing such synthesized sound signals, a voice recognition system can perform voice recognition with a high recognition rate/low misrecognition rate, a cellular phone and a transceiver allow conversations in high-quality sound, and a broadcasting system allows high-quality broadcasting and recording.

As in the foregoing first embodiment, the use of the modal inclination $\tau 0$ in the estimation of the delay time also allows robust voice input, even in the case of environmental changes, such as a change in the mounted positions of the microphones, and movement of the sound source, such as movement or a change in posture of the speaker. In other words, robust voice input can be achieved while keeping a high degree of

freedom in the positions of the microphones.

As described above, the aforementioned various effects can be attained even on the assumption of the use of mountable microphones, which are compact and lightweight for easy mounting/removing, can ensure short-distance sound pickup generally as good as close-talking microphones, and can ease the burden and discomfort when mounted to the user compared to close-talking microphone headsets.

(Example (Second Embodiment))

A voice recognition experiment with acoustic models was conducted using the synchronized added sound (synthesized sound signals) generated by a system to which the present invention was applied.

In this voice recognition experiment with acoustic models, first of all, acoustic models were prepared using learning data obtained from the synchronized added sound. The acoustic models prepared were as follows:

(1) Four collection-environment-dependent HMMs (hidden Markov models) prepared for each collection environment, and

(2) a collection-environment-independent HMM acquired through learning using sound from all the collection environments.

The collection environments were the same as above: in a soundproof room, in a car, in an office space, and on a sidewalk.

Then, a voice recognition experiment was conducted using the prepared acoustic models.

The recognition task was continuous voice recognition, and the data for evaluation (sound for evaluation) were different sound from that used in the learning. FIG. 10 shows the results of the voice recognition experiment. The results of the recognition rate with the mono-channel sound from the first and second microphones 1 and 2 are also shown as comparison objects (conventional examples). The first and second microphones 1 and 2 were a glasses microphone and a chest microphone, respectively, for example. The glasses microphone refers to a microphone mounted to the frame of glasses.

As shown in FIG. 10, the result was that the recognition rate with the synchronized added sound obtained by the present invention exceeded the recognition rate with the mono-channel sound in a

soundproof room, on a sidewalk, and in all the environments, except in a car. This demonstrated that the synchronized added sound generated by the system to which the present invention was applied was of high quality also in actual environments.

A third embodiment is described hereinafter.

FIG. 11 shows the construction of a sound signal processor 10 of the third embodiment. The sound signal processor 10 of the second embodiment is a combined form of the sound signal processors 10 of the foregoing first and second embodiments. That is, the sound signal processor 10 of the third embodiment includes a voiced/voiceless determining section 34, a delay processing section 51, a waveform synthesizing section 52, and a sound input on/off control section 18 at the same time.

Constructed as above, the sound signal processor 10 of the third embodiment operates as follows. Note that those sections not specifically described operate in the same manner as in the sound signal processors 10 of the foregoing first and second embodiments.

The delay processing section 51 delays the sound signals of the second microphone 2 based on the modal inclination $\tau 0$ calculated by the histogram etc. calculating section 33 (mode calculating section 33C). The waveform synthesizing section 52 synthesizes the sound signals of the second microphone 2 inputted from and delayed by the delay processing section 51 and the sound signals from the first microphone 1, and outputs the synthesized sound signals to the sound input on/off control section 18.

Meanwhile, the voiced/voiceless determining section 34 determines voiced and voiceless sections based on the incidence obtained by the histogram etc. calculating section 33, and the sound input on/off control section 18 switches on and off to and not to output the sound signals (synchronized added sound signals) outputted from the waveform synthesizing section 52 based on the determination results.

Constructed as above, the sound signal processor 10 of the third embodiment can demonstrate the effects achieved by the sound signal processors 10 of the foregoing first and second embodiments.

That is, high-quality synthesized sound signals can be generated,

allowing accurate detection of speech sections contained therein. Further, even in the case of environmental changes, such as a change in the mounted positions of the microphones, and movement of the sound source, such as movement or a change in posture of the speaker, robust voice input can be achieved. In other words, robust voice input can be achieved while keeping a high degree of freedom in the positions of the microphones.

The descriptions of the embodiments of the present invention have been made above. The application of the present invention, however, is not limited to the foregoing embodiments.

For example, as shown in FIG. 12, the voiced/voiceless determining section 34 compares the inclinations τi calculated by the first through N-th inclination calculating sections 32_1 through 32_N and the modal inclination $\tau 0$, using the following inequality (9):

$$|\tau i - \tau 0| < \alpha\sigma \qquad (9)$$

wherein α represents a coefficient, and σ represents a value physically included within the threshold used for the determination (inclination threshold) δ described previously. For example, the point of providing δ and $\alpha\sigma$ is to distinguish the difference between the effects in detecting voiced sections due to the both values, namely δ as a constant and $\alpha\sigma$ as a variable progressively updated through real-time learning.

Since σ in $\alpha\sigma$ is updatable, the conditions for the determination of a voiced section may be made more strict to more securely prevent misdetermination of a voiceless section in quiet environments. Meanwhile, the conditions for the determination may be made less strict to allow more stable detection of a voiced section in environments with background noise. Assuming that σ adapted for quiet environments is used in environments with background noise, which case is equivalent to the case when δ as a constant is used, there is a fear that a voiced section carrying overlapped noise and voice may be missed.

In other words, δ as a constant functions effectively in the detection of voiced sections when used in environments similar to the conditions under which that value was set, while $\alpha\sigma$ as a variable functions effectively in the detection of voiced sections when used in a system

intended to dynamically respond to environmental changes.

The strictness of the determination may be increased and reduced by changing the coefficient α.

In the foregoing embodiments, the tendency that the inclinations of each band concentrate on a specific inclination was observed by creating histograms from these inclinations of each band. However, the tendency that the inclinations of each band concentrate on a specific inclination may be observed by another method.

Also, in the descriptions of the foregoing embodiments, the detection target sound was speech sound produced by humans. However, the detection target sound may be sound produced by substances other than humans.

In the descriptions of the foregoing embodiments, the first and second framing sections 11 and 12, first and second frequency analyzing sections 13 and 14, and cross-spectrum calculating section 15 implement cross-spectrum phase detection means for detecting the phase of a cross-spectrum between the sound signals inputted into plural microphones; the phase extracting section 16, phase unwrap processing section 17, frequency band dividing section 31, and first through N-th inclination calculating sections 32_1 through 32_N implement inclination detection means for detecting the inclination of the phase of the cross-spectrum detected by the cross-spectrum phase detection means with respect to the frequency; and the histogram etc. calculating section 33 and voiced/voiceless determining section 34 implement speech sound detection means for detecting a speech section contained in the sound picked up by the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means.

In addition, the histogram etc. calculating section 33 and delay processing section 51 implement delay time detection means for detecting the delay time between the sound signals picked up by the plural microphones based on the inclination with respect to the frequency detected by the inclination detection means; and the waveform synthesizing section 52 implements sound signal synthesizing means for synthesizing the sound signals inputted into the plural

microphones based on the delay time detected by the delay time detection means.

Further, the sound signal processor 10 of the foregoing embodiments may be applied to a voice recognition device. In this case, the voice recognition device includes voice recognition processing means for performing voice recognition processing of the sound signals contained in the speech section (speech sound) detected by the sound signal processor 10, in addition to the components of the sound signal processor 10 as described above.

Examples of voice recognition techniques include "VORERO" (trademark), a voice recognition technique proposed by Asahi Kasei Kabushiki Kaisha (see http://www.asahi-kasei.co.jp/vorero/jp/vorero/feature.html). The present invention may be applied to voice recognition devices using such voice recognition techniques.

Furthermore, the sound signal processor 10 of the foregoing embodiments may be implemented on a computer. And, the processing operation of the sound signal processor 10 as described above may be performed on a computer with a predetermined program. In this case, such a program may be designed to make the computer perform processing of detecting a target sound, the processing including: inputting detection target sounds outputted from a detection target sound source into plural microphones; detecting the phase of a cross-spectrum between the sound signals inputted into the plural microphones; detecting the inclination of the phase of the cross-spectrum with respect to the frequency due to the respective distances from the detection target sound source to the plural microphones; and, based on the inclination, detecting the target sound outputted from the detection target sound source and picked up by the plural microphones. Alternatively, the program may be designed to make the computer perform processing of detecting the delay time in sound input, the processing including: inputting sounds outputted from a sound source into plural microphones; detecting the phase of a cross-spectrum between the sound signals inputted into the plural microphones; detecting the inclination of the phase of the cross-spectrum with respect to the

frequency due to the respective distances from the sound source to the plural microphones; and, based on the inclination, detecting the delay time in sound reception from the sound source between the plural microphones.

Industrial Applicability

The present invention allows construction of a sound reception system employing mountable microphones and robust against environmental fluctuations.